**National Quality Forum—Measure Testing (subcriteria 2a2, 2b2-2b6)**

**Measure Title**: Glycemic Control – Hypoglycemia

**Date of Submission**: 12/5/2013

**Type of Measure:**

|  |  |
| --- | --- |
| ☐ Composite – ***STOP – use composite testing form*** | Outcome (*including PRO-PM*) |
| ☐ Cost/resource | ☐ Process |
| ☐ Efficiency | ☐ Structure |

|  |
| --- |
| **Instructions**   * Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form. * **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.** * **For outcome and resource use measures**, section **2b4** also must be completed. * If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed. * Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed. * If you are unable to check a box, please highlight or shade the box for your response. * Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.*** * Contact NQF staff regarding questions. Check for resources at [Submitting Standards webpage](http://www.qualityforum.org/Measuring_Performance/Submitting_Standards.aspx). |

|  |
| --- |
| **Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF’s evaluation criteria for testing.**  **2a2.** **Reliability testing** [**10**](#Note10) demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise.  **2b2.** **Validity testing** [**11**](#Note11) demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality.    **2b3.** Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [**12**](#Note12)  **AND**  If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [**13**](#Note13)  **2b4.** **For outcome measures and other measures when indicated** (e.g., resource use):   * **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [**14**](#Note14)**,**[**15**](#Note15) and has demonstrated adequate discrimination and calibration   **OR**   * rationale/data support no risk adjustment/ stratification.   **2b5.** Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [**16**](#Note16) **differences in performance**;  **OR**  there is evidence of overall less-than-optimal performance.  **2b6.** **If multiple data sources/methods are specified, there is demonstration they produce comparable results**.  **Notes**  **10.** Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).  **11.** Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.  **12.** Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.  **13.** Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.  **14.** Risk factors that influence outcomes should not be specified as exclusions.  **15.** Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.  **16.** With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers. |

**1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE**

*Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.*

**1.1. What type of data was used for testing**? (*Check all the sources of data identified in the measure specifications and data used for testing the measure*. *Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

|  |  |
| --- | --- |
| **Measure Specified to Use Data From:**  **(*must be consistent with data sources entered in S.23*)** | **Measure Tested with Data From:** |
| ☐ abstracted from paper record | ☐ abstracted from paper record |
| ☐ administrative claims | ☐ administrative claims |
| ☐ clinical database/registry | ☐ clinical database/registry |
| abstracted from electronic health record | abstracted from electronic health record |
| eMeasure (HQMF) implemented in EHRs | eMeasure (HQMF) implemented in EHRs |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

**1.2. If an existing dataset was used, identify the specific dataset** (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

Not applicable

**1.3. What are the dates of the data used in testing**? Admissions from January 3, 2011 – December 31, 2012

**1.4. What levels of analysis** **were tested**? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

|  |  |
| --- | --- |
| **Measure Specified to Measure Performance of:**  **(*must be consistent with levels entered in item S.26*)** | **Measure Tested at Level of:** |
| ☐ individual clinician | ☐ individual clinician |
| ☐ group/practice | ☐ group/practice |
| hospital/facility/agency | hospital/facility/agency |
| ☐ health plan | ☐ health plan |
| ☐ other: Click here to describe | ☐ other: Click here to describe |

**1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

A total of eight hospitals, comprising seven acute care hospitals and one critical access hospital (CAH), from four states (AZ, FL, MO, and TX) were used to perform the field testing of the measure. FMQAI included a variety of hospital sites within its field testing bed to encompass urban and rural, large and small, teaching and non-teaching, and for-profit and non-profit types. In addition, FMQAI included hospitals with EHR systems with the highest market share, such as Epic and Cerner, as well as a hospital with a home-grown, hybrid version of EHR systems that qualified for Stage I Meaningful Use. Table 1 provides a breakdown of the characteristics of the participating hospitals included in the field testing of the measure.

**Table 1. Field Testing Hospital Characteristics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Hospital ID** | **Location** | **Size** | **Type** | **Teaching Status** | **Ownership** | **EHR System** |
| 1 | Urban | 852 | Level 1 Trauma | Teaching | Non-Profit | Epic |
| 2 | Urban | 522 | Level 1 Trauma | Teaching | Non-Profit | Epic |
| 3 | Rural | 025 | Critical Access Hospital | Non-Teaching | Non-Profit | Cerner |
| 4 | Urban | 312 | Acute Regional Hospital | Teaching | For Profit | McKesson |
| 5 | Urban | 493 | Level 1 Trauma | Non-Teaching | For Profit | Cerner |
| 6 | Urban | 359 | Acute Care Community Hospital | Teaching | For Profit | Cerner |
| 7 | Urban | 143 | Acute Care Community Hospital | Teaching | For Profit | Cerner |
| 8 | Urban | 695 | Level 1 Trauma | Teaching | Non-Profit | Epic |

**1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)**? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*). EHR data extracts obtained from the 8 participating hospitals in 4 states (AZ, FL, MO, and TX) were used to support the field testing of the measure. The criterion for the field testing sample was to include approximately 5,000 inpatient admissions or at least 1 year of admissions per hospital. Table 2 shows the demographic characteristics of the sample by hospital.

**Table 2. Demographic Characteristics of the Field Testing Sample**

| **Demographic** | **Hospital 1** | **Hospital 2** | **Hospital 3** | **Hospital 4** | **Hospital 5** | **Hospital 6** | **Hospital 7** | **Hospital 8** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gender |  |  |  |  |  |  |  |  |
| Female | 4,531 (55.05%) | 3,095 (51.51%) | No data | 5,563 (50.80%) | 4,320 (49.57%) | 5,792 (68.59%) | 4,722 (54.89%) | 3,005 (55.80%) |
| Male | 3,700 (44.95%) | 2,914 (48.49%) | No data | 5,388 (49.20%) | 4,395 (50.43%) | 2,652 (31.41%) | 3,880 (45.11%) | 2,380 (44.20%) |
| Age |  |  |  |  |  |  |  |  |
| ≥ 65 years | 2,257 (27.42%) | 571 (9.5%) | 153 (68.00%) | 8,615 (78.67%) | 5,988 (68.71%) | 3,215 (38.07%) | 4,380 (50.92%) | 1,102 (20.46%) |
| Race |  |  |  |  |  |  |  |  |
| White/Caucasian | 5,658 (68.74%) | 2,338 (38.91%) | 184 (81.78%) | 10,762 (98.72%) | 7,785 (89.33%) | 7,933 (93.95%) | 7,592 (88.26%) | 2,499 (44.36%) |
| African-American | 1,792 (21.77%) | 733 (12.20%) | 17 (7.56%) | 51 (0.47%) | 615 (7.06%) | 133 (1.58%) | 798 (9.28%) | 2,814 (49.95%) |
| Hispanic | 353 (4.29%) | 2,502 (41.64%) | 0 | 32 (0.30%) | 0 | 0 | 0 | 0 |
| Other | 428 (5.20%) | 436 (7.25%) | 24 (10.66%) | 106 (0.97%) | 315 (3.61%) | 378 (4.47%) | 212 (2.46%) | 321 (5.69%) |
| Ethnicity |  |  |  |  |  |  |  |  |
| Hispanic | 309 (3.75%) | 2,318 (38.58%) | 27 (12.00%) | 16 (0.15%) | 290 (3.33%) | 6,445 (76.33%) | 52 (0.60%) | 237 (4.21%) |
| Non-Hispanic | 7,921 (96.25%) | 3,691 (61.42%) | 198 (88.00%) | 10,935 (99.85%) | 8,425 (96.67%) | 1,999 (23.67%) | 8,550 (99.40%) | 5,397 (95.79%) |
| Payor Source |  |  |  |  |  |  |  |  |
| Medicare | 3,100 (37.66%) | 926 (15.41%) | 185 (82.22%) | 8,698 (79.43%) | 6,241 (71.61%) | 3,571 (42.29%) | 5,229 (60.79%) | 1,777 (33.00%) |
| Medicaid | 1,682 (20.43%) | 0 | 12 (5.33) | 292 (2.67%) | 367 (4.21%) | 1,848 (21.89%) | 406 (4.72%) | 1,927 (35.78%) |
| Self-Pay | 951 (11.55%) | 2,029 (33.77%) | 6 (2.67%) | 336 (3.07%) | 492 (5.65%) | 631 (7.47%) | 187 (2.17%) | 50 (0.93%) |
| Other | 2,498 (30.35%) | 3,054 (50.82%) | 22 (9.78%) | 1,625 (14.84%) | 1,615 (18.53%) | 2,394 (28.35%) | 2,780 (32.32%) | 1,631 (30.29%) |

**1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below**.

Subsamples were randomly selected from the testing sample described in Section 1.6 for testing the criterion and construct validity of the measure.

Criterion (Content) Validity

The subsample for the criterion validity testing consisted of a random selection of 50 inpatient admissions from each of the 8 participating hospitals (n=400), and 398 of the 400 cases were reviewed by a trained registered nurse abstractor. Of these cases, 121 (30%) were selected for re-review by another trained registered nurse abstractor.

Construct Validity

The subsample for the construct validity testing consisted of a random selection of 107 inpatient encounters from the overall inpatient encounters received from the 8 participating hospitals. All 107 cases were reviewed by trained physician reviewers. Of these cases, 26 (24%) were selected for re-review by another trained physician reviewer.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2a2. RELIABILITY TESTING**

***Note****: If accuracy/correctness (validity) of data elements was empirically tested*, *separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter “see section 2b2 for validity testing of data elements”; and skip 2a2.3 and 2a2.4.*

**2a2.1. What level of reliability testing was conducted**? (*may be one or both levels*)  
 **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)   
 **Performance measure score** (e.g., *signal-to-noise analysis*)   
  
**2a2.2. For each level checked above, describe the method of reliability testing and what it tests** (*describe the steps―do not just name a method; what type of error does it test; what statistical analysis was used*)

The accuracy/correctness (validity) of data elements for this measure was empirically tested. Therefore, per the NQF Measure Evaluation Criteria and Guidance, separate reliability testing of data elements was not required. Please refer to the Validity Testing section for additional testing information on the scientific acceptability of the measure properties.

In order to assess measure precision in the context of the observed variability across measurement units (hospital facilities), FMQAI utilized the approach proposed by Adams (2009). The rationale for this choice of testing was based on the work on the reliability for provider profiling for the National Committee for Quality Assurance (NCQA). The following is quoted from the tutorial published by Adams: “Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing the number of patients in the physician’s data as well as increasing the number of measures per patient.”

The signal-to-noise ratio was calculated as a function of the variance between hospitals (signal) and the variance within a hospital (noise). Reliability was estimated using a beta-binomial model. This approach has two basic assumptions:

1. Each hospital has a true pass rate, p, which varies from hospital to hospital; and
2. The hospital’s score is a binomial random variable conditional on the hospital’s true value, which comes from the beta distribution.

Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual hospital variance); whereas, a reliability of 1.0 implies that all variation is caused by a real difference in performance (across hospitals). In a simulation, Adams showed that differences between providers started to be detectable at reliability of 0.7, and significant differences could be seen at reliability of 0.9. Our rationale was based on Adams’ work; thus, a minimum reliability score of 0.7 was used to indicate sufficient signal strength to discriminate performance between hospitals.

Citation

Adams, J. L. The reliability of provider profiling: A tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009.

**2a2.3. For each level checked above, what were the statistical results from reliability testing**? (e*.g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)  
Reliability tests were conducted across hospitals, and the results are displayed in Table 3.

**Table 3. Reliability Statistics by Hospitals**

|  |  |  |  |
| --- | --- | --- | --- |
| **Hospital** | **Denominator** | **Measure Rate** | **Reliability Score** |
| **1** | 11,530 | 0.57% | 0.77 |
| **2** | 06,149 | 0.36% | 0.70 |
| **3** | 00340 | 0.88% | 0.08 |
| **4** | 11,939 | 0.67% | 0.75 |
| **5** | 11,827 | 0.57% | 0.77 |
| **6** | 09,812 | 0.89% | 0.67 |
| **7** | 13,316 | 0.57% | 0.79 |
| **8** | 07,045 | 0.41% | 0.71 |

**2a2.4 What is your interpretation of the results in terms of demonstrating reliability**? (i*.e., what do the results mean and what are the norms for the test conducted?*)  
Six hospitals have scores that are greater than or equal to 0.70, and one hospital has a score approaching 0.70. The reliability score for Hospital 3 was significantly lower than the other hospitals. Hospital 3 is a small-size critical access hospital with 25 beds, yielding a much smaller denominator and a limited number of hypoglycemic cases (n=3). The reliability test showed that the measure rates are reliable across the large and medium-size hospitals.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b2. VALIDITY TESTING**

**2b2.1. What level of validity testing was conducted**? (*may be one or both levels*)  
 **Critical data elements** (*data element validity must address ALL critical data elements*)

**Performance measure score**

☐ **Empirical validity testing** **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

**2b2.2. For each level checked above, describe the method of validity testing and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*  
Criterion (Content) Validity

Criterion validity evaluated whether the measure correctly identified all the data elements that were required to calculate the measure rate. In addition, this analytic method quantified the agreement between the extraction of data obtained directly from the participating providers’ EHR systems using the QDM logic applied to the data elements obtained by review and abstraction of the entire EHR as a gold standard. The criterion chart review was performed by registered nurses trained by FMQAI, using a standardized chart review electronic abstraction tool, developed in Microsoft Access. Nurse abstractors were instructed to verify that the received data elements that were required to calculate the measure rate from the data extraction matched the information contained within the EHR.

Construct Validity

Construct validity addressed whether the measure identified true manifestation of hypoglycemic events as defined by the measure specifications. In addition, the construct validity testing assessed whether additional causes could be attributed to the hypoglycemic event to inform risk factors that would need to be considered for risk stratification purposes. The construct validity chart review was conducted by physician reviewers trained by FMQAI and the University of Florida, using a standardized chart review paper abstraction tool.

Using the above-described sample of charts, the physician reviewers were asked to assess construct validity using the following quantitative and qualitative questions:

1. Did you find solid evidence that the event really occurred? (yes/no)
2. Were there any treatments given to resolve the hypoglycemic events? (Such as D50, orange juice, etc.)
3. Did you identify additional causes that can explain the occurrence of hypoglycemia? If so, do you suggest specific variables that should be included in risk adjustment algorithms if hospitals were to be compared?

Face Validity

FMQAI’s Technical Expert Panel (TEP) evaluated the face validity of the measure and measure score after field testing was completed. The names and organizations of TEP members are listed in Table 4. The evaluation of face validity was conducted through an online review process using a web-based questionnaire (developed using SurveyMonkey®). TEP members were specifically asked “whether the performance score from the measure as specified represents an accurate reflection of quality of care,” and responded by indicating their level of agreement with the statement on a 5-point Likert scale (1=Strongly Disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly Agree).

**Table 4. TEP Members**

| **Name** | **Organization** |
| --- | --- |
| **Dale W. Bratzler**  DO, MPH | Professor and Associate Dean, College of Public Health, University of Oklahoma Health Sciences Center |
| **Mary Brennan-Taylor** | Adjunct Research Instructor of Family Medicine, School of Medicine and Biomedical Sciences, University of Buffalo  Representing: TEP as Patient Representative |
| **Frank E. Briggs III** PharmD, MPH | Vice President, Quality and Patient Safety, West Virginia University Healthcare  Representing: American Society of Health-System Pharmacists |
| **Daniel Castillo**  MD, MBA | Medical Director, Healthcare Quality Evaluation, The Joint Commission |
| **Joan Ching**  RN, MN, CPHQ | Administrative Director, Hospital Quality & Safety, Virginia Mason Medical Center |
| **Edward S. Eisenberg**  MD, FACP | Senior Vice President, Performance Measurement and Strategic Alliances, Pharmacy Quality Alliance |
| **Floyd Eisenberg**  MD, MPH, FACP | President, iParsimony, LLC |
| **Marybeth Farquhar**  PhD, MSN, RN | Vice President of Research & Measurement, URAC |
| **Frank Federico**  BS, RPh | Executive Director for Strategic Partners, Institute for Healthcare Improvement |
| **Robert Feroli**  PharmD, FASHP | Medication Safety Officer, Johns Hopkins Hospital |
| **Tejal Gandhi**  MD, MPH | President, National Patient Safety Foundation; Board-certified Internist and Associate Professor of Medicine, Harvard Medical School  Representing: American Hospital Association |
| **P. Michael Ho**  MD, PhD, FACC | Staff Cardiologist, VA Eastern Colorado Health Care System; Associate Professor of Medicine, University of Colorado Denver  Representing: American College of Cardiology |
| **Mark L. Holtsman** PharmD | Co-Director, Inpatient Pain Service and Pain Management Service, Pharmacist, UC Davis Medical Center; Clinical Professor of Anesthesiology and Pain Medicine, UC Davis School of Medicine  Representing: American Academy of Pain Medicine |
| **Clifford Ko**  MD, MS, MSHS, FACS | Director, ACS Division of Research and Optimal Patient Care; Director, ACS NSQIP; Professor of Surgery and Health Services, UCLA Schools of Medicine and Public Health  Representing: American College of Surgeons |
| **Janet Maurer**  MD, MBA, FCCP | Operations Medical Director, National Imaging Associates, Health Dialog; Clinical Professor of Medicine, University of Arizona, College of Medicine, Phoenix Campus; Staff Physician, St. Joseph’s Medical Center  Representing: American College of Chest Physicians |
| **Michael N. Neuss**  MD | Chief Medical Officer, Vanderbilt-Ingram Cancer Center; Professor of Medicine, Vanderbilt School of Medicine  Representing: American Society of Clinical Oncology |
| **N. Lee Rucker**  MSPH | Senior Advisor, National Council on Patient Information and Education |
| **Edward Septimus**  MD, FACP, FIDSA, FSHEA | Medical Director, Infection Prevention and Epidemiology Clinical Service Group, HCA Healthcare System; Clinical Professor of Internal Medicine, Texas A&M University  Representing: Infectious Diseases Society of America |
| **Nathan Spell**  MD, FACP | Chief Quality Officer, Emory University Hospital; Associate Professor of Medicine, Emory University School of Medicine  Representing: American College of Physicians |
| **Stephen J. Traub**  MD, FACEP | Assistant Professor in Emergency Medicine and Chair, Department of Emergency Medicine, Mayo Clinic  Representing: American College of Emergency Physicians |
| **Darren M. Triller**  PharmD | Senior Director, Quality Improvement, IPRO QIO |

**2b2.3. What were the statistical results from validity testing**? (*e.g., correlation; t-test*)   
Criterion (Content) Validity

Criterion validity was summarized as the percent of data elements that were correctly identified by the electronically extracted measure, according to the chart review. As noted above, inter-rater reliability, or the proportion of agreement concerning the accuracy of the data elements received from the data extract, was calculated using the 398 charts that were reviewed by the nurse abstractors. The proportion of agreement on each data element is shown in Table 5 by hospital.

Overall, 98.97% of the data elements found in the medical record correctly matched the EHR data extract received from the participating hospitals. The data element with the highest criterion validity score was the “Discharge\_DateTime” at 100% and the data element with the lowest criterion validity score was the “Medication Administered: Anti-diabetic DateTime” at 97.71%.

**Table 5. Results of Criterion Validity Testing**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Element** | **Hospital 1** | **Hospital 2** | **Hospital 3** | **Hospital 4** | **Hospital 5** | **Hospital 6** | **Hospital 7** | **Hospital 8** | **Average** |
| Date\_of\_Birth | 100% | 100% | 93.75% | 100% | 100% | 100% | 100% | 100% | 99.22% |
| Admission\_DateTime | 96.00% | 100% | N/A | 100% | 100% | 100% | 100% | 100% | 99.43% |
| Discharge\_DateTime | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |
| Medication Administered: Anti-diabetic | 98.00% | 100% | 100% | 98.00% | 100% | 100% | 98.00% | 100% | 99.25% |
| Medication Administered:  Anti-diabetic DateTime | 98.00% | 96.00% | 91.67% | 98.00% | 100% | 100% | 98.00% | 100% | 97.71% |
| Lab\_Test: BG < 40 | 98.00% | 100% | 100% | 100% | 96.00% | 98.00% | 96.00% | 100% | 98.50% |
| Total Proportion Agreement | 98.33% | 99.33% | 97.08% | 99.33% | 99.33% | 99.67% | 98.67% | 100% | 98.97% |

One hundred and twenty-one cases were re-reviewed by a different nurse abstractor to check the reliability of the abstraction. Of these hypoglycemic cases reviewed by both nurse abstractors, the inter-rater reliability (proportion of agreement) score was 97.9%. Since this value was above the 95% minimum acceptable score, no further re-reviews were required. All discrepancies found were analyzed and discussed by both abstractors as an additional training tool to aid in quality improvement for future abstractions.

Construct Validity

Construct validity was summarized as the percent of cases detected by the measure that were true cases of manifest hypoglycemic events, and the test results showed that 95.3% of the 107 cases reviewed had a true hypoglycemic manifestation confirmed by the physician reviewers. The measure was modified as a result of the 5 cases (4.7%) that were found to be false positive manifestation events. For the 26 hypoglycemic cases reviewed by both physician reviewers, there was 100% agreement between reviewers.

Face Validity

Eighteen of the 21 (86%) TEP members completed the face validity evaluation for the measure. The results of the TEP rating of face validity on a scale of 1 to 5 are presented in Table 6.

**Table 6. Results of the Face Validity Evaluation**

|  |  |
| --- | --- |
| **Rating** | **Number of TEP (%)** |
| 5 (Strongly Agree) | 8 (44.4%) |
| 4 (Agree) | 10 (55.6%) |
| 3 (Neutral) | 0 |
| 2 (Disagree) | 0 |
| 1 (Strongly Disagree) | 0 |

All TEP members agreed that the measure was valid as specified. The mean rate was 4.44, and the median rate was 4.

**2b2.4. What is your interpretation of the results in terms of demonstrating validity**? (i*.e., what do the results mean and what are the norms for the test conducted?*)  
The results from content validity testing demonstrated that the data elements in the data extract generated from the EHR systems are accurate when compared to manual abstraction of the full medical record from the EHR. The results from the construct validity assessment indicate that the measure accurately captures inpatient hypoglycemia as defined by the specifications. Furthermore, all 18 of the TEP members that completed the evaluation either responded “agree” or “strongly agree” that the measure, as specified, exhibited face validity.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b3. EXCLUSIONS ANALYSIS**

☐ **no exclusions — *skip to section*** [***2b4***](#section2b4)

**2b3.1. Describe the method of testing exclusions and what it tests** (*describe the steps―do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

The TEP recommended that the measure exclude admissions with lengths of stay greater than 120 days. The exclusion is established for the measure to be consistent with measures in the CMS Hospital Inpatient Quality Reporting Program. In addition, the follow-up period is restricted to avoid artificial inflation of the measure rate due to multiple hypoglycemic events from a small number of admissions with extended lengths of stay. The number affected by this exclusion was examined, and the measure rates with and without the exclusion were calculated and compared.

**2b3.2. What were the statistical results from testing exclusions**? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

**Table 7. Exclusion of Qualified Denominator Admissions with Lengths of Stay Greater than 120 Days**

|  |  |  |  |
| --- | --- | --- | --- |
| **Hospital** | **Number Excluded (%)** | **Measure Rate With Exclusion** | **Measure Rate Without Exclusion** |
| Hospital 5 | 1 (0.05%) | 0.57% | 0.57% |

The other seven field testing hospitals did not have any admissions with lengths of stay greater than 120 days that were excluded from the measure denominator.

**2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results?** (*i.e., the value outweighs the burden of increased data collection and analysis.*  *Note:* ***If patient preference is an exclusion****, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

There was only one admission that was excluded from the measure denominator due to the length of stay greater than 120 days. The exclusion did not impact the measure rate; however, the exclusion was maintained to align the proposed measure with other measures in the CMS Hospital Inpatient Quality Reporting Program.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES**  
***If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section*** [***2b5***](#section2b5)***.***

**2b4.1. What method of controlling for differences in case mix is used?**

**No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**

☐ **Other,** Click here to enter description

**2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities**.

Hypoglycemia, even though it can pose the potential for patient harm, such as in cases of hypoglycemic coma, is considered a surrogate (or intermediate) outcome. As with process measures, where differences in patient case mix can determine whether a particular best practice (process) is harder or easier to implement, achievement of acceptable surrogate outcomes thresholds may be affected in a similar fashion. However, reasoning for omitting risk adjustment algorithms for process and surrogate outcomes is based on the observation that the process can be implemented for most patients (or the surrogate outcome value achieved) and that a limited number of causes exist that may inhibit such achievement.

Causes for hypoglycemia are quite limited, including primarily either a lack of caloric intake or overuse of anti-diabetic agents. As both are well controllable, especially in hospital environments, limited unpreventable causes for hypoglycemia exist.

Chart review during the formative testing phase identified two scenarios where hypoglycemia might not be preventable. The first included end of life situations when a patient is provided with comfort measures only. A sensitivity analysis examining the incidence of hypoglycemia in the patients at the formative testing hospital suggested that these events were too rare to affect measure rates. A second scenario involved the treatment of cardiac arrest or hyperkalemia, where often doses of regular insulin are added. Even though these doses are usually combined with dextrose, clinical reviewers suggested that individual responses to insulin in insulin-naïve patients are not always fully predictable and may require larger doses of dextrose than expected. This also was deemed to be a rare event with limited impact on measure rates.

It was noted that the underlying need for glucose management could affect the risk for hypoglycemia. For example, oral anti-diabetic agents are less likely to cause hypoglycemia, whereas patients with the need for insulin would have a higher risk. However, the selected threshold for hypoglycemia at <40mg/dL was considered sufficiently conservative to expect that such events should have been prevented regardless of the extent of glucose management that is required.

**2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk** (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

Not applicable

**2b4.4. What were the statistical results of the analyses used to select risk factors?**

Not applicable

**2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach** (*describe the steps―do not just name a method; what statistical analysis was used*)  
Not applicable

*Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below*.  
***if stratified, skip to*** [***2b4.9***](#question2b49)

**2b4.6. Statistical Risk Model Discrimination Statistics** (*e.g., c-statistic, R-squared*)**:**

**2b4.7. Statistical Risk Model Calibration Statistics** (*e.g., Hosmer-Lemeshow statistic*):

**2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves**:

**2b4.9. Results of Risk Stratification Analysis**:

**2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)?** (i*.e., what do the results mean and what are the norms for the test conducted*)

\***2b4.11.** **Optional Additional Testing for Risk Adjustment** (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods*)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE**

**2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified** (*describe the steps―do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

To identify statistically significant differences in performance, chi-square tests of homogeneity were conducted to detect differences at the hospital level. If significant differences were identified, pairwise comparisons were then used to identify differences between specific hospitals.

**2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities?** (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)  
Table 8 shows the measure rates by hospital, the 95% confidence intervals, and the mean measure rate, which indicates the average measure performance across the hospitals.

**Table 8. Meaningful Differences in Measure Rates**

|  |  |  |  |
| --- | --- | --- | --- |
| **Hospital ID** | **Measure Rate** | **95% Confidence Interval** | **Statistical Difference Hospital** |
| **1** | 0.57% | 0.44%-0.73% | 2 and 6 (p-value ≤0.035) |
| **2** | 0.36% | 0.22%-0.54% | 1, 4, 5, 6, 7 (p-value ≤ 0.003) |
| **3** | 0.88% | 0.18%-2.56% | No Statistical Difference identified |
| **4** | 0.67% | 0.53%-0.83% | 2, 6 and 8 (p-value ≤ 0.0002) |
| **5** | 0.57% | 0.44%-0.73% | 2, 6 and 8 (p-value ≤ 0.049) |
| **6** | 0.89% | 0.71%-1.09% | 1, 2, 5, 7 & 8 (p-value ≤ 0.0002) |
| **7** | 0.57% | 0.45%-0.71% | 2, 6, & 8 (p-value ≤ 0.049) |
| **8** | 0.41% | 0.28%-0.59% | 4, 5, 6, & 7 (p-value ≤ 0.049) |
| **Mean** | **0.62%** | N/A | N/A |

N/A = Not applicable

Table 9 shows hospitals with statically significant differences in the measure rates including p-values. Hospital 3 measure rate is no different from any of the other hospitals’ rates. The remaining seven hospitals differ statistically with at least two other facilities.

**Table 9. P-values for Pairwise Comparisons of Measure Rates**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Hospital ID**  Measure Rate | **1**  0.57% | **2**  0.36% | **3**  0.88% | **4**  0.67% | **5**  0.57% | **6**  0.89% | **7**  0.57% | **8**  0.41% |
| **1**  0.57% |  | 0.0035 |  |  |  | <0.0001 |  |  |
| **2**  0.36% | 0.0035 |  |  | <0.0001 | 0.0029 | <0.0001 | 0.0027 |  |
| **3**  0.88% |  |  |  |  |  |  |  |  |
| **4**  0.67% |  | <0.0001 |  |  |  | 0.0002 |  | <0.0001 |
| **5**  0.57% |  | 0.0029 |  |  |  | <0.0001 |  | 0.0489 |
| **6**  0.89% | <0.0001 | <0.0001 |  | 0.0002 | <0.0001 |  | <0.0001 | <0.0001 |
| **7**  0.57% |  | 0.0027 |  |  |  | <0.0001 |  | 0.0490 |
| **8**  0.41% |  |  |  | <0.0001 | 0.0489 | <0.0001 | 0.0490 |  |

**2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities?** (*i.e., what do the results mean in terms of statistical and meaningful differences?*)  
The result demonstrated that statistically significant differences can be detected between hospitals. The variations in performance among the hospitals suggested meaningful differences in the quality of care provided between the lowest and highest performing hospital and indicated that there is room for improvement.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS**

***If only one set of specifications, this section can be skipped.***

**Note***: This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

**2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different datasources/specifications** (*describe the steps―do not just name a method; what statistical analysis was used*)  
Not applicable

**2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications?** (*e.g., correlation, rank order*)

Not applicable

**2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications?** (i*.e., what do the results mean and what are the norms for the test conducted*)  
Not applicable